

WHAT IS CLAIMED IS:

1. An apparatus for processing a binary floating-point number having a sign bit and a mantissa having a fraction portion comprising:

a fraction mask table configured to identify the fraction portion of the binary floating-point number; and

a multiplexer configured to replace each bit of the fraction portion with the sign bit, thereby producing a floor of the binary floating-point number.

2. The apparatus of claim 1, further comprising:

a decrementer configured to decrement the binary floating-point number before replacing when the binary floating-point number is negative.

3. The apparatus of claim 2, further comprising:

a format converter configured to convert the floor to two's complement format.

4. The apparatus of claim 3, wherein the format converter comprises:

an exclusive-OR gate configured to perform an exclusive-OR operation between each bit of the floor and the sign bit, thereby producing a result of the exclusive-OR operation; and

means for concatenating the sign bit and the result of the exclusive-OR operation, thereby producing a signed two's complement mantissa of the floor.

5. The apparatus of claim 4, wherein the format converter further comprises:

a signed-right-shift shifter configured to perform, upon the signed two's complement mantissa of the floor, a signed-right-shift operation, thereby producing the floor of the binary floating-point number in two's complement format.

6. The apparatus of claim 1, further comprising:

a format converter configured to convert the floor to floating-point format.

7. The apparatus of claim 6, wherein the format converter comprises:

an incrementer configured to increment the floor when the binary floating-point number is negative, and doing nothing otherwise, thereby producing an incremented value; and

means for replacing the most-significant bit (MSB) of the incremented value with the exponent bits and the sign bit, such that the sign bit is the MSB, thereby producing a floor of the binary floating-point number in floating-point format.

8. The apparatus of claim 7, wherein the binary floating-point number includes an exponent that differs from an unbiased exponent by a bias offset, and wherein the incrementer is configured to increment the floor when the binary floating-point number is negative and the exponent is greater than, or equal to, the bias offset, thereby producing an incremented value.

9. The apparatus of claim 8, wherein the binary floating-point number includes an exponent that differs from an unbiased exponent by a bias offset, and wherein the incrementer is further configured to:

replace the exponent bits with the offset when the binary floating-point number is negative and the exponent is less than the offset; and

replace the exponent bits with zeros when the binary floating-point number is positive and the exponent is less than the offset.

10. The apparatus of claim 1, further comprising a floating-point subtractor configured to take a floating-point difference between the binary floating-point number and the floor of the binary floating-point number, thereby producing a fractional remainder of the binary floating-point number.

11. The apparatus of claim 1, wherein the binary floating-point number includes an exponent that differs from an unbiased exponent by a bias offset, and wherein the multiplexer is further configured to replace each bit of the fraction portion with zero ("0") when the exponent is greater than, or equal to, the bias offset.

12. The apparatus of claim 2, wherein the binary floating-point number includes an exponent that differs from an unbiased exponent by a bias offset, and wherein the decrementer is further configured to decrement the binary floating-point number before processing by the multiplexer when the binary floating-point number is negative unless the exponent is less than the bias offset.

13. A apparatus for determining the floating-point floor of a floating-point number, the apparatus comprising:

means for identifying a binary floating-point number including a sign bit, exponent bits, and mantissa bits, wherein the binary floating-point number is negative when the sign bit is a one ("1");

means for concatenating an implicit bit and the mantissa bits, thereby producing a first binary number such that the implicit bit is the most significant bit (MSB) of the first binary number;

a decrementer configured to decrement the first binary number when the sign bit is a one ("1") and do nothing when the sign bit is a zero ("0"), thereby producing a second binary number;

a fraction mask table configured to identify a fraction portion of the second binary number based upon a predetermined exponent bias;

a multiplexer configured to replace each bit of the fraction portion with the sign bit thereby producing a third binary number;

an exclusive-OR configured to perform an exclusive-OR operation between each bit of the third binary number and the sign bit thereby producing a fourth binary number;

means for concatenating the sign bit, the fourth binary number, and a first predetermined number of zeros, thereby producing a fifth binary number such that the sign bit is the MSB of the fifth binary number, and the zeros are the least significant bits of the fifth binary number; and

a signed-right-shift configured to perform, upon the fifth binary number, a signed-right-shift operation by a number of bits equivalent to the difference between the exponent and a second predetermined number, thereby producing the floor of the binary floating-point number in integer format.

14. The apparatus of claim 13, further comprising:

means for subtracting one from the sum of the number of bits in the floating-point floor and the number of bits in the third binary number, thereby producing the first predetermined number.

15. The apparatus of claim 13, wherein the exponent differs from an unbiased exponent of the floating-point number by a bias offset, further comprising:

means for summing the bias offset, the number of bits in the mantissa, and the first predetermined number, thereby producing the second predetermined number.

16. The apparatus of claim 13, wherein the multiplexer is further configured to: accept a fraction mask corresponding to the exponent bits, the fraction mask having a one for each bit belonging to the fraction portion and a zero for each bit belonging to the integer portion; and

apply the fraction mask to the second binary number.

17. The apparatus of claim 13, wherein the implicit bit is zero ("0") when the exponent bits are all zero ("0") and the implicit bit is one ("1") otherwise.

18. The apparatus of claim 13, wherein the binary floating-point number includes an exponent that differs from an unbiased exponent of the floating-point number by a bias offset, and wherein the multiplexer is further configured to:

replace each bit of the fraction portion with the sign bit when the exponent is greater than, or equal to, the bias offset, thereby producing a third binary number.

19. The apparatus of claim 13, wherein the binary floating-point number includes an exponent that differs from an unbiased exponent of the floating-point number by a bias offset, and wherein the decrementer is further configured to:

decrement the first binary number when the sign bit is a one ("1") and doing nothing when the sign bit is a zero ("0"), unless the exponent is less than the bias offset, thereby producing a second binary number.

5 20. The apparatus of claim 13, further comprising:

 means for replacing the MSB of the second binary number with the exponent bits and the sign bit, thereby producing a sixth binary number, such that the sign bit is the MSB of the sixth binary number;

 means for replacing the MSB of the third binary number with the exponent bits and
10 the sign bit, thereby producing a seventh binary number, such that the sign bit is the MSB of the seventh binary number; and

 a floating-point subtractor configured to perform a floating-point subtraction with the sixth binary number as the minuend and the seventh binary number as the subtrahend, thereby producing a fractional remainder of the floating-point number.

15 21. A method for processing a binary floating-point number having a sign bit and a mantissa having a fraction portion comprising:

 identifying the fraction portion of the binary floating-point number; and

 replacing each bit of the fraction portion with the sign bit, thereby producing a floor
20 of the binary floating-point number.

 22. The method of claim 21, further comprising:

 decrementing the binary floating-point number before replacing when the binary floating-point number is negative.

25 23. The method of claim 22, further comprising:

 converting the floor to two's complement format.

 24. The method of claim 23, wherein converting comprises:

30 performing an exclusive-OR operation between each bit of the floor and the sign bit, thereby producing a result of the exclusive-OR operation; and

concatenating the sign bit and the result of the exclusive-OR operation, thereby producing a signed two's complement mantissa of the floor.

25. The method of claim 24, wherein converting further comprises:
5 performing, upon the signed two's complement mantissa of the floor, a signed-right-shift operation, thereby producing the floor of the binary floating-point number in two's complement format.

26. The method of claim 1, wherein converting comprises:21, further comprising:
10 converting the floor to floating-point format.

27. The method of claim 26, wherein converting comprises:
incrementing the floor when the binary floating-point number is negative, and doing
nothing otherwise, thereby producing an incremented value; and
15 replacing the most-significant bit (MSB) of the incremented value with the exponent bits and the sign bit, such that the sign bit is the MSB, thereby producing a floor of the binary floating-point number in floating-point format.

28. The method of claim 27, wherein the binary floating-point number includes an
20 exponent that differs from an unbiased exponent by a bias offset, and wherein incrementing comprises:

incrementing the floor when the binary floating-point number is negative and the
exponent is greater than, or equal to, the bias offset, thereby producing an incremented value.

25 29. The method of claim 28, wherein the binary floating-point number includes an exponent that differs from an unbiased exponent by a bias offset, and wherein incrementing further comprises:

replacing the exponent bits with the offset when the binary floating-point number is
negative and the exponent is less than the offset; and

30 replacing the exponent bits with zeros when the binary floating-point number is positive and the exponent is less than the offset.

30. The method of claim 21, further comprising
taking a floating-point difference between a value of the binary floating-point number
before replacing and a value of the binary floating-point number after replacing, thereby
5 producing a fractional remainder of the binary floating-point number.

31. The method of claim 21, wherein the binary floating-point number includes an
exponent that differs from an unbiased exponent by a bias offset, and wherein replacing
comprises:

10 replacing each bit of the fraction portion with zero ("0") when the exponent is greater
than, or equal to, the bias offset.

32. The method of claim 22, wherein the binary floating-point number includes an
exponent that differs from an unbiased exponent by a bias offset, and wherein decrementing
15 comprises:

decrementing the binary floating-point number before replacing when the binary
floating-point number is negative unless the exponent is less than the bias offset.

33. A method for determining the floating-point floor of a floating-point number,
the method comprising:

20 identifying a binary floating-point number including a sign bit, exponent bits, and
mantissa bits, wherein the binary floating-point number is negative when the sign bit is a one
("1");

25 concatenating an implicit bit and the mantissa bits, thereby producing a first binary
number such that the implicit bit is the most significant bit (MSB) of the first binary number;

decrementing the first binary number when the sign bit is a one ("1") and doing
nothing when the sign bit is a zero ("0"), thereby producing a second binary number;

identifying a fraction portion of the second binary number based upon a
predetermined exponent bias;

30 replacing each bit of the fraction portion with the sign bit thereby producing a third
binary number;

performing an exclusive-OR operation between each bit of the third binary number and the sign bit thereby producing a fourth binary number;

concatenating the sign bit, the fourth binary number, and a first predetermined number of zeros, thereby producing a fifth binary number such that the sign bit is the MSB of the fifth binary number, and the zeros are the least significant bits of the fifth binary number; and

performing, upon the fifth binary number, a signed-right-shift operation by a number of bits equivalent to the difference between the exponent and a second predetermined number, thereby producing the floor of the binary floating-point number in integer format.

34. The method of claim 33, further comprising:

subtracting one from the sum of the number of bits in the floating-point floor and the number of bits in the third binary number, thereby producing the first predetermined number.

35. The method of claim 33, wherein the exponent differs from an unbiased exponent of the floating-point number by a bias offset, further comprising:

summing the bias offset, the number of bits in the mantissa, and the first predetermined number, thereby producing the second predetermined number.

36. The method of claim 33, wherein replacing comprises:

identifying a fraction mask corresponding to the exponent bits, the fraction mask having a one for each bit belonging to the fraction portion and a zero for each bit belonging to the integer portion; and

applying the fraction mask to the second binary number.

37. The method of claim 33, wherein the implicit bit is zero ("0") when the exponent bits are all zero ("0") and the implicit bit is one ("1") otherwise.

38. The method of claim 33, wherein the binary floating-point number includes an exponent that differs from an unbiased exponent of the floating-point number by a bias offset, and wherein replacing comprises:

replacing each bit of the fraction portion with the sign bit when the exponent is greater than, or equal to, the bias offset, thereby producing a third binary number.

39. The method of claim 33, wherein the binary floating-point number includes an exponent that differs from an unbiased exponent of the floating-point number by a bias offset, and wherein decrementing comprises:

decrementing the first binary number when the sign bit is a one ("1") and doing nothing when the sign bit is a zero ("0"), unless the exponent is less than the bias offset, thereby producing a second binary number.

40. The method of claim, further comprising:

replacing the MSB of the second binary number with the exponent bits and the sign bit, thereby producing a sixth binary number, such that the sign bit is the MSB of the sixth binary number;

replacing the MSB of the third binary number with the exponent bits and the sign bit, thereby producing a seventh binary number, such that the sign bit is the MSB of the seventh binary number; and

performing a floating point subtraction with the sixth binary number as the minuend and the seventh binary number as the subtrahend, thereby producing a fractional remainder of the floating-point number.

41. A computer program product, tangibly stored on a computer-readable medium, for processing a binary floating-point number having a sign bit and a mantissa having a fraction portion, comprising instructions operable to cause a programmable processor to:

identify the fraction portion of the binary floating-point number; and

replace each bit of the fraction portion with the sign bit, thereby producing a floor of the binary floating-point number.

42. The computer program product of claim 41, further comprising instructions operable to cause a programmable processor to:

decrement the binary floating-point number before replacing when the binary floating-point number is negative.

43. The computer program product of claim 42, further comprising instructions operable to cause a programmable processor to:
convert the floor to two's complement format.

44. The computer program product of claim 43, wherein instructions operable to cause a programmable processor to convert comprise instructions operable to cause a programmable processor to:

perform an exclusive-OR operation between each bit of the floor and the sign bit, thereby producing a result of the exclusive-OR operation; and
concatenate the sign bit and the result of the exclusive-OR operation, thereby producing a signed two's complement mantissa of the floor.

45. The computer program product of claim 44, wherein instructions operable to cause a programmable processor to convert further comprise instructions operable to cause a programmable processor to:

perform, upon the signed two's complement mantissa of the floor, a signed-right-shift operation, thereby producing the floor of the binary floating-point number in two's complement format.

46. The computer program product of claim 41, further comprising instructions operable to cause a programmable processor to:
convert the floor to floating-point format.

47. The computer program product of claim 46, wherein instructions operable to cause a programmable processor to convert comprise instructions operable to cause a programmable processor to:

increment the floor when the binary floating-point number is negative, and doing nothing otherwise, thereby producing an incremented value; and

replace the most-significant bit (MSB) of the incremented value with the exponent bits and the sign bit, such that the sign bit is the MSB, thereby producing a floor of the binary floating-point number in floating-point format.

5 48. The computer program product of claim 47, wherein the binary floating-point number includes an exponent that differs from an unbiased exponent by a bias offset, and wherein instructions operable to cause a programmable processor to increment comprise instructions operable to cause a programmable processor to:

10 increment the floor when the binary floating-point number is negative and the exponent is greater than, or equal to, the bias offset, thereby producing an incremented value.

15 49. The computer program product of claim 48, wherein the binary floating-point number includes an exponent that differs from an unbiased exponent by a bias offset, and wherein instructions operable to cause a programmable processor to increment further comprise instructions operable to cause a programmable processor to:

 replace the exponent bits with the offset when the binary floating-point number is negative and the exponent is less than the offset; and

20 replace the exponent bits with zeros when the binary floating-point number is positive and the exponent is less than the offset.

25 50. The computer program product of claim 41, further comprising instructions operable to cause a programmable processor to:

 take a floating-point difference between a value of the binary floating-point number before replacing and a value of the binary floating-point number after replacing, thereby producing a fractional remainder of the binary floating-point number.

30 51. The computer program product of claim 41, wherein the binary floating-point number includes an exponent that differs from an unbiased exponent by a bias offset, and wherein instructions operable to cause a programmable processor to replace comprise instructions operable to cause a programmable processor to:

replace each bit of the fraction portion with zero ("0") when the exponent is greater than, or equal to, the bias offset.

52. The computer program product of claim 42, wherein the binary floating-point number includes an exponent that differs from an unbiased exponent by a bias offset, and wherein instructions operable to cause a programmable processor to decrement comprise instructions operable to cause a programmable processor to:

decrement the binary floating-point number before replacing when the binary floating-point number is negative unless the exponent is less than the bias offset.

53. A computer program product, tangibly stored on a computer-readable medium, for determining the floating-point floor of a floating-point number, comprising instructions operable to cause a programmable processor to:

identify a binary floating-point number including a sign bit, exponent bits, and mantissa bits, wherein the binary floating-point number is negative when the sign bit is a one ("1");

concatenate an implicit bit and the mantissa bits, thereby producing a first binary number such that the implicit bit is the most significant bit (MSB) of the first binary number;

decrement the first binary number when the sign bit is a one ("1") and doing nothing when the sign bit is a zero ("0"), thereby producing a second binary number;

identify a fraction portion of the second binary number based upon a predetermined exponent bias;

replace each bit of the fraction portion with the sign bit thereby producing a third binary number;

perform an exclusive-OR operation between each bit of the third binary number and the sign bit thereby producing a fourth binary number;

concatenate the sign bit, the fourth binary number, and a first predetermined number of zeros, thereby producing a fifth binary number such that the sign bit is the MSB of the fifth binary number, and the zeros are the least significant bits of the fifth binary number; and

perform, upon the fifth binary number, a signed-right-shift operation by a number of bits equivalent to the difference between the exponent and a second predetermined number, thereby producing the floor of the binary floating-point number in integer format.

5 54. The computer program product of claim 53, further comprising:
 subtracting one from the sum of the number of bits in the floating-point floor and the
 number of bits in the third binary number, thereby producing the first predetermined number.

10 55. The computer program product of claim 53, wherein the exponent differs from
 an unbiased exponent of the floating-point number by a bias offset, further comprising
 instructions operable to cause a programmable processor to:

 sum the bias offset, the number of bits in the mantissa, and the first predetermined
 number, thereby producing the second predetermined number.

15 56. The computer program product of claim 53, wherein instructions operable to
 cause a programmable processor to replace comprise instructions operable to cause a
 programmable processor to:

 identify a fraction mask corresponding to the exponent bits, the fraction mask having
 a one for each bit belonging to the fraction portion and a zero for each bit belonging to the
20 integer portion; and

 apply the fraction mask to the second binary number.

25 57. The computer program product of claim 53, wherein the implicit bit is zero
 ("0") when the exponent bits are all zero ("0") and the implicit bit is one ("1") otherwise.

 58. The computer program product of claim 53, wherein the binary floating-point
 number includes an exponent that differs from an unbiased exponent of the floating-point
 number by a bias offset, and wherein instructions operable to cause a programmable
 processor to replace comprise instructions operable to cause a programmable processor to:

30 replace each bit of the fraction portion with the sign bit when the exponent is greater
 than, or equal to, the bias offset, thereby producing a third binary number.

59. The computer program product of claim 53, wherein the binary floating-point number includes an exponent that differs from an unbiased exponent of the floating-point number by a bias offset, and wherein instructions operable to cause a programmable processor to decrement comprise instructions operable to cause a programmable processor to:
decrement the first binary number when the sign bit is a one ("1") and do nothing when the sign bit is a zero ("0"), unless the exponent is less than the bias offset, thereby producing a second binary number.

60. The computer program product of claim 53, further comprising instructions operable to cause a programmable processor to:

replace the MSB of the second binary number with the exponent bits and the sign bit, thereby producing a sixth binary number, such that the sign bit is the MSB of the sixth binary number;

replace the MSB of the third binary number with the exponent bits and the sign bit, thereby producing a seventh binary number, such that the sign bit is the MSB of the seventh binary number; and

perform a floating point subtraction with the sixth binary number as the minuend and the seventh binary number as the subtrahend, thereby producing a fractional remainder of the floating-point number.